



**I Международная
научно-практическая
конференция
ЯДЕРНЫЕ ЗНАНИЯ В XXI ВЕКЕ
Сборник научных трудов**

**26 ноября 2025 года
Минск, Беларусь**



ПОЛНОТЕКСТОВЫЙ ПОИСК НА НАЦИОНАЛЬНОМ ПОРТАЛЕ ЯДЕРНЫХ ЗНАНИЙ РЕСПУБЛИКИ БЕЛАРУСЬ

Статья посвящена краткому обзору реализации полнотекстового поиска на Национальном портале ядерных знаний Республики Беларусь.

Рассмотрим ведущие мировые порталы в области ядерных знаний – “Атомная энергия 2.0” и портал МАГАТЭ. На российском портале в области ядерных знаний, работающем под эгидой Росатома, “Атомная энергия 2.0” <https://www.atomic-energy.ru/>, который позиционирует себя как «семантическое научное СМИ» (<https://www.atomic-energy.ru/video/119454>), семантические технологии реализуются вручную: «Научный портал “Атомная энергия 2.0” с момента создания в 2008 году развивается как открытая семантическая информационная система по управлению ядерными знаниями и взаимодействию с широкой общественностью. Сегодня на портале накоплено уже более 120 000 публикаций, каждая из которых вручную “семантически” отсортирована по общей тематике (1 000+ терминов), ключевым словам (1 500+ терминов), географии (1 000+ терминов), организациям (2 000+ наименований), персоналиям (1 500+ наименований) и событиям (1 000+ наименований)».

Портал “Атомная энергия 2.0” написан на системе управления контентом (Content Management System – CMS) Drupal, в которой используемый встроенный поиск не способен искать по части слова (только по полному совпадению слов), а также по присоединенным документам и т.д. Разработчики, использующие CMS Drupal, вынуждены использовать встроенный поиск Google или Яндекс.

На портале МАГАТЭ <https://www.iaea.org/> используется встроенный поиск Google.

На Национальном портале ядерных знаний Республики Беларусь реализован собственный полнотекстовый поиск, в том числе по части слов (леммам), по всему содержимому записи – заголовку, авторам, выходным данным, резюме, самому материалу, а также по присоединенным файлам, включая сканированные документы и изображения.

Ни на одном из известных общепризнанных порталов ядерных знаний (“Атомная энергия 2.0”, МАГАТЭ и др.) нет опции полнотекстового поиска по сканированным документам и изображениям.

Приведем обоснование необходимости собственной встроенной системы полнотекстового поиска на Национальном портале.

1. Собственный поиск в системе работает при отсутствии подключения к интернет и не требует доступа к внешним серверам. Россия активно строит локальный сегмент глобальной сети, в результате чего периодически становятся недоступны сервера Яндекса и другие сервисы. Google и Microsoft Bing явля-

ются компаниями из недружественных стран и в любой момент их услуги могут стать недоступны.

2. Информация, размещенная на Национальном портале, не передается на сторонние серверы. Таким образом можно лучше контролировать доступ к данным.

3. Возможность локально хранить языковую модель и использовать ее для развития системы необходима, например, для построения классификатора документов.

4. Ликвидация эффекта «отравления данных» поисковых систем важна, поскольку корпорации могут ранжировать результаты в соответствии со своими различными бизнес-интересами (заказы рекламодателей, действующие судебные решения или предъявленные иски, политические предпочтения владельцев бизнеса и т.п.). Сейчас ядерные технологии вернули себе позицию в топе глобальной повестки, поэтому независимость в подаче данных очень важна.

5. Возможность поддерживать при необходимости сторонние источники данных может быть реализована именно через собственный поиск, который можно оснастить средствами для обработки сторонних сайтов, если будет принято и организационно обеспечено соответствующее решение.

Реализованный полнотекстовый поиск на Национальном портале является семантическим, основанным на извлечении текста из файлов посредством Apache Tika, получению лемм по алгоритму Snowball для контекстного понимания содержания, модели Okapi BM25 – функции ранжирования, используемой для упорядочивания документов по их релевантности данному поисковому запросу. Полнотекстовый поиск в настоящий момент ведется по всем записям на портале и присоединенным документам – то есть по всему контенту портала.

Итак, информация на портале размещается, как правило, в виде текстов и этой информации много. Возникает задача поиска текстов с учетом морфологии (полнотекстовый поиск) как для авторов, так и для посетителей сайта. Собственная подсистема полнотекстового поиска портала обладает следующими ключевыми возможностями:

- поиск с учетом морфологии естественного языка;
- ранжирование тестов по релевантности;
- поддержка большого числа форматов документов, включая возможность извлечения текста из отсканированных PDF-документов и изображений (обеспечивается программным пакетом Apache Tika в комплексе с ПО Tesseract).
- тезаурус (система глоссариев) для организации структуры публикуемых текстов.
- подсистема полуавтоматического поиска и расстановки ключевых слов для документов.

Работа поисковой системы организована следующим образом. Проводится декомпозиция стадий обработки и выделение в фоновые задачи тех из них, которые невозможно выполнить при прямом взаимодействии пользователя с системой. Кроме этого, обработка отсканированных PDF, может быть выполнена только в фоновом режиме. Обычно страница текста на типичном хостинге распознается в течение одной-двух минут, а документ может содержать несколько десятков страниц. Т. е. оптимальной является работа в фоновом режиме.

Так же фоновый режим удобен для решения следующих задач:

- анализ морфологии текста с использованием словарей и эвристических алгоритмов;
- выделение ключевых слов и словосочетаний;
- автоматическое формирование индексов (структур, оптимизирующих поиск).

В итоге большинство стадий анализа текста – это фоновая обработка отдельным фоновым сервисом (демоном). Данный процесс запускается автоматически при старте системы.

Пользовательский интерфейс позволяет загрузить данные и отображает промежуточные результаты в те моменты, когда нужна реакция пользователя, и, конечно, показывает итоговые результаты.

Процесс взаимодействия с фоновой обработкой осуществляется путем построения взаимодействия через базу данных (БД). Современные системы управления базами данных (СУБД) обладают всеми необходимыми для этого возможностями и позволяют решить множество проблем своими штатными средствами, начиная от совместного доступа к данным до контроля целостности с применением транзакций. СУБД Firebird, используемая в системе управления контентом портала, позволила оптимизировать взаимодействие и избежать периодического опроса БД фоновым процессом, что снизило нагрузку на процессор. Это достигается путем применения специального инструмента, доступного в СУБД Firebird, такого как события (Firebird Events).

Событие срабатывает, когда пользователь размещает в системе новый файл или редактирует/удаляет старый. Информация об этом практически мгновенно передается в фоновый процесс и активирует обработку, что позволяет избежать задержек, характерных для редкого опроса и неоправданного расхода ресурсов процессора при частом опросе. Фоновая обработка может занять очень разное время от 2–3 секунд до десятков минут. При этом возможна комфортная работа пользователей над другими задачами.

Результаты обработки хранятся в БД системы в виде двух наборов данных:

- результаты извлечения чистого текста из документов для морфологического анализа;
- расчет частоты вхождения слов в текст с учетом морфологии текста на русском и английском языках.

Полученные данные записываются в БД, а когда возникает необходимость, они извлекаются из БД и применяются в расчетах.

Так же необходимо отметить, что современные семантические алгоритмы очень требовательны к ресурсам компьютера, поэтому становится актуальной задача оптимизировать их выполнение. Отметим следующие способы автоматизации, которые применены при реализации системы.

Частоты вхождения слов в текст – это самый часто применяемый набор данных в системе, который используется каждым ее модулем. Для ускорения обработки запросов они записываются в двух форматах:

- обычные нормализованные таблицы, которые предоставляют возможность удобного вычисления статистики, используя SQL;
- поле в одной из таблиц с денормализованными данными, что позволяет полу-

чить вектор частот слов всего текста одним запросом в БД без формирования выборки.

В системе реализовано кэширование редко изменяющихся наборов данных. Так же неочевидной, но полезной особенностью оказалось хранение чистого текста документа в БД, реализованное в виде полезной функции предпросмотра текста документов.

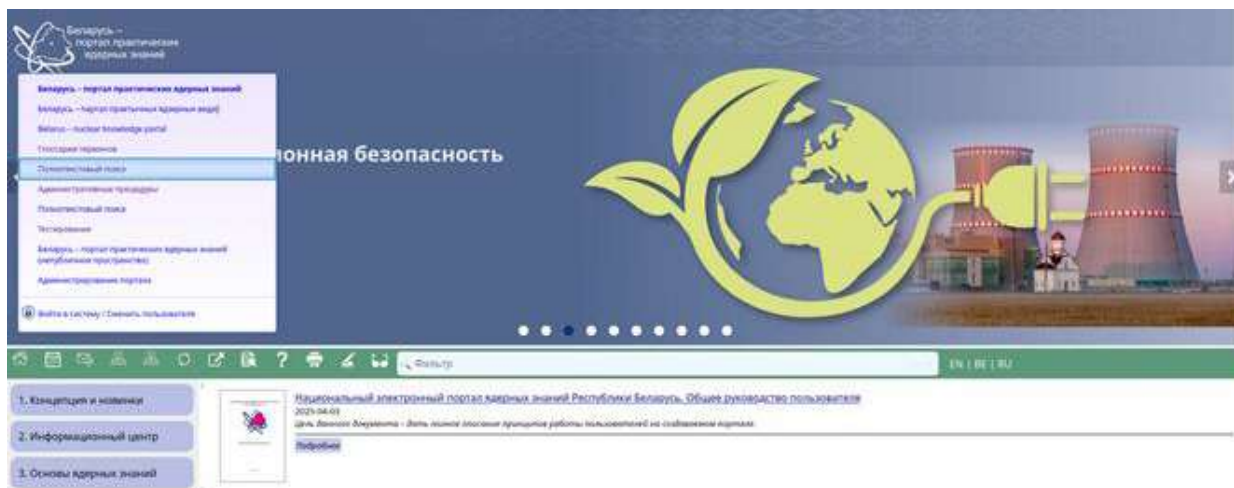


Рисунок 1. Меню для перехода в режим полнотекстового поиска

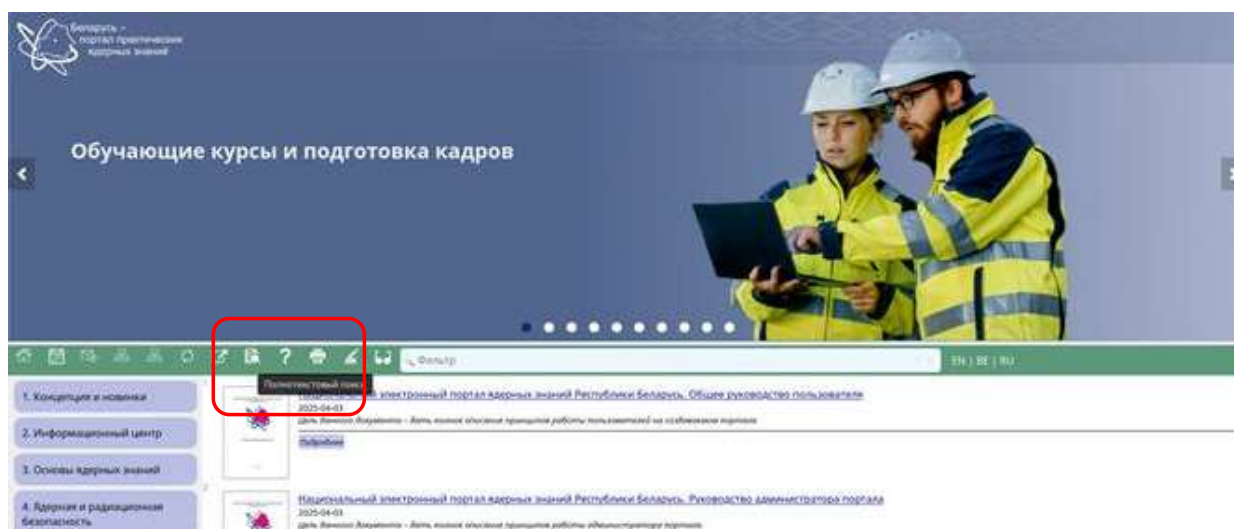


Рисунок 2. Меню для перехода в режим полнотекстового поиска

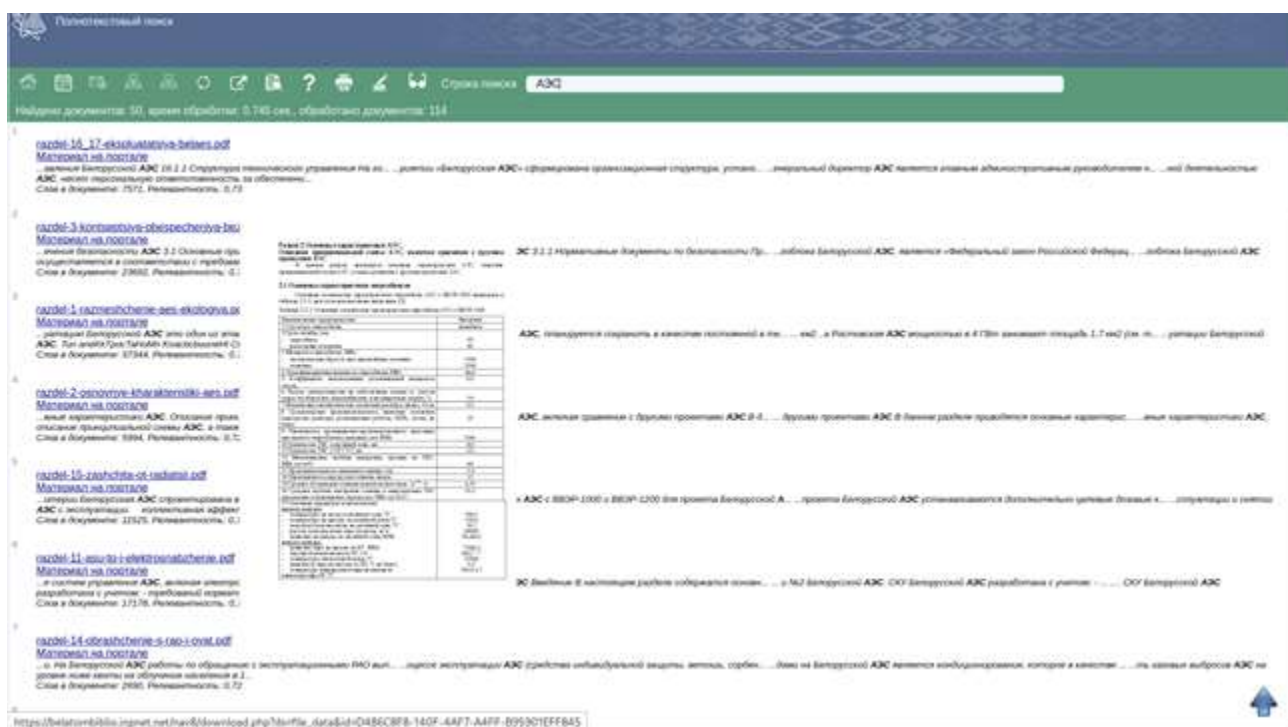


Рисунок 3. Интерфейс пользователя в режиме полнотекстового поиска

Рассмотрим, как в системе (<https://belatombiblio.inpnet.net/>) реализован полнотекстовый поиск по текстовым документам (включая картинки и сканированные документы), которые загружены в систему. Для перехода в режим полнотекстового поиска необходимо выбрать строку «Полнотекстовый поиск» в меню в левом верхнем углу экрана модуля представления (рис.1) либо на панели навигации (рис. 2).

После перехода системы на экран полнотекстового поиска (рис. 3) необходимо набрать текстовый запрос в строке поиска в левой верхней части экрана. В данном случае набрана поисковая строка «АЭС».

Результат выполнения поиска высвечивается на странице в виде имени документа (документов) либо материалов (записей) портала, где найдена поисковая строка. Здесь выделены несколько слов текста из поисковой строки. В системе доступен предпросмотр страницы текста документа и открытие в новом окне самого документа (рис. 5).

Также в верхней части экрана выводится количество документов, в которых обнаружена строка ввода. Найденные документы и материалы высвечиваются на странице в порядке релевантности (от наибольшего значения — числа вхождений поисковой строки в документ, до наименьшего).

Предпросмотр страницы текста документа, присоединенного к записи или рисунка доступен также при работе с материалами (вложенными файлами) на панели «Подробности» (рис. 32) при наведении курсора мыши на название файла.

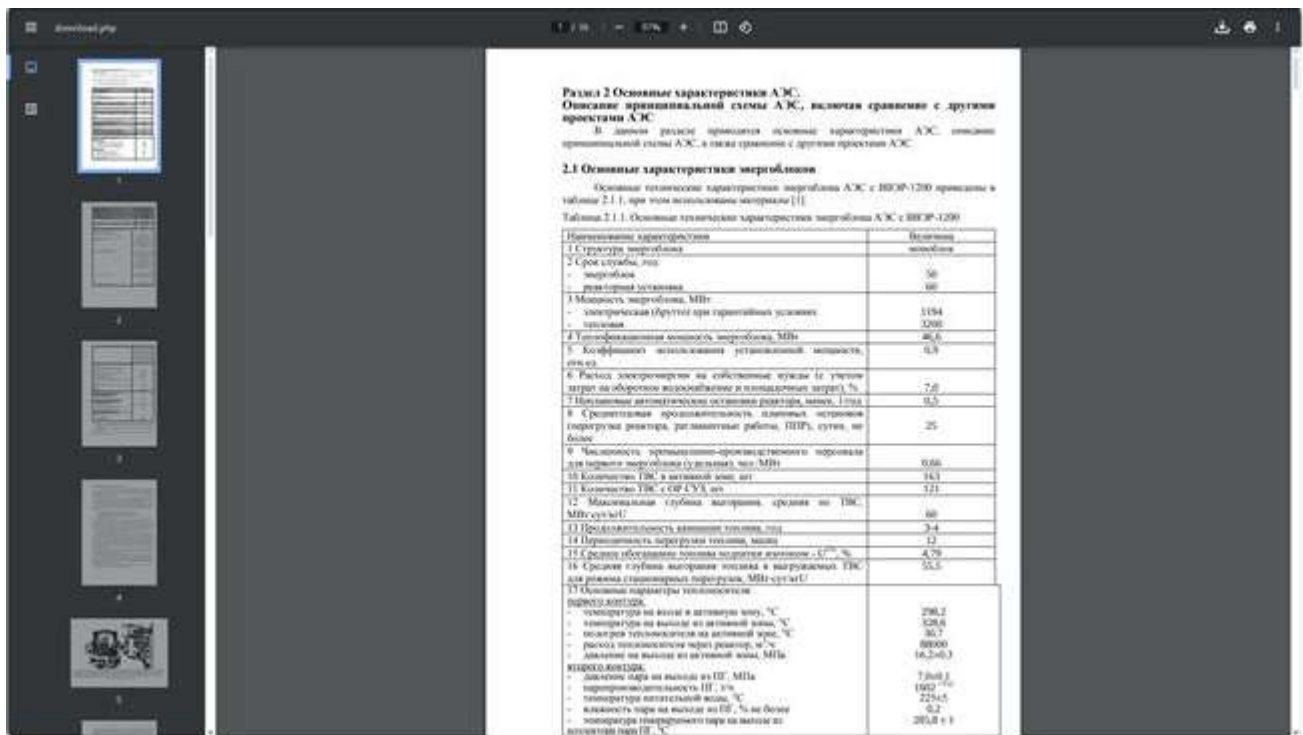


Рисунок 4. Интерфейс пользователя в режиме полнотекстового поиска.
 Просмотр найденного документа

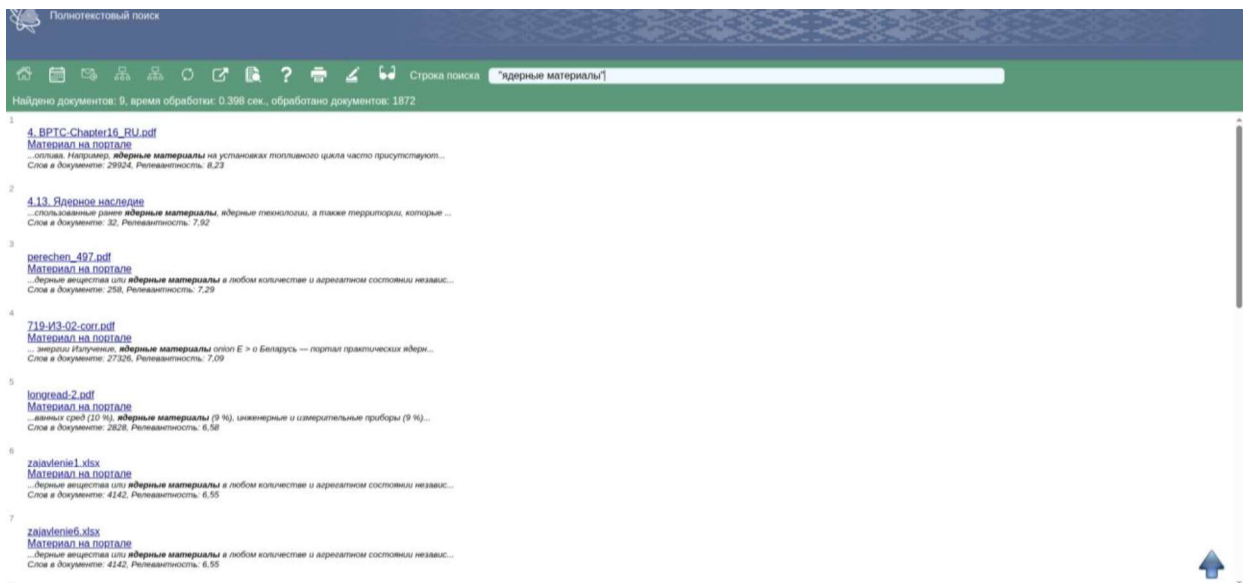


Рисунок 5. Интерфейс пользователя в режиме полнотекстового поиска. Полное совпадение фразы

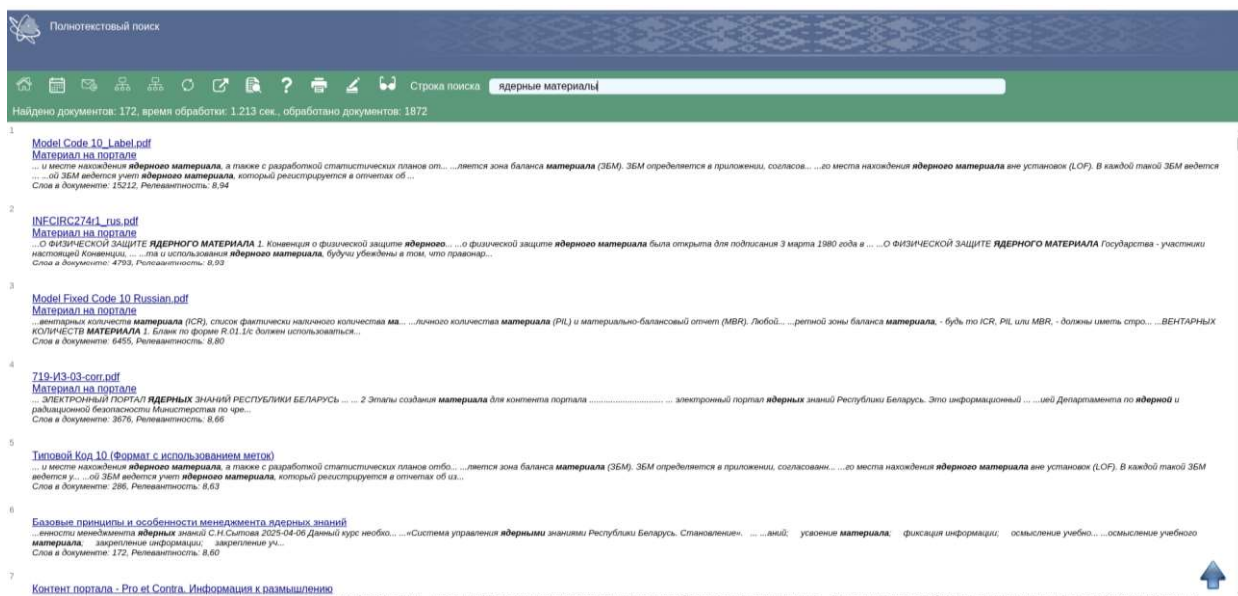


Рисунок 6. Интерфейс пользователя в режиме полнотекстового поиска.
Функция «и» для всех словоформ

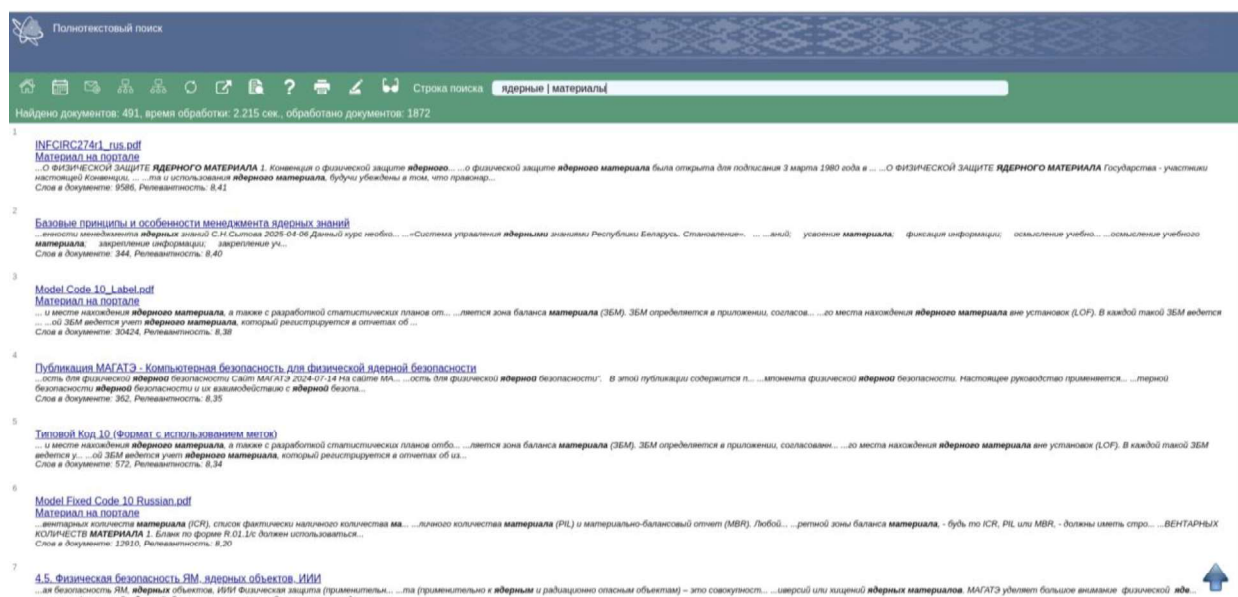


Рисунок 7. – Полнотекстовый поиск. Функция «или» для всех словоформ

Для поиска фразы целиком (по полному совпадению) фраза должна заключаться в кавычки: **“ядерные материалы”** (рис. 5).

Фраза без кавычек **ядерные материалы** реализует поиск по вхождению всех возможных словоформ всех слов фразы, возможно, в произвольном порядке (функция «и») (рис. 6).

Справочно: словоформы слова «материалы» – материал, материала, материалу, материалом, материале, материалов и т.д.

Фраза **ядерные | материалы** реализует поиск всех словоформ входящих во фразу слов, одного или нескольких (функция «или») (рис. 7).

Функция «или» реализована через знак | с целью возможности использовать слово «или» в поиске наравне с другими. Так же может быть использовано ||.

В системе доступно распознавание не только офисных документов и документов типа pdf, но и сканированных документов, а также картинок в форматах jpg, tiff и др.

Отметим, что в процессе встроенного распознавания исходный документ не изменяется и не осуществляется ручное исправление ошибок. Пользователь, при желании, может скопировать документ и распознать его дополнительно с помощью сторонней опции. Достаточное разрешение для гарантированного распознавания документа – 300 dpi. Если в сканированном документе крупный шрифт (14 pt и больше), то можно использовать разрешение от 100 dpi.

Работы выполняются в рамках мероприятия 13 «Выполнение работ по оказанию научно-технической поддержки Министерству по чрезвычайным ситуациям Республики Беларусь в области обеспечения ядерной и радиационной безопасности» подпрограммы 3 «Научное обеспечение эффективной и безопасной работы Белорусской атомной электростанции и перспективных направлений развития атомной энергетики» Государственной программы «Научные технологии и техника» на 2021–2025 годы.